AUDIO FORENSICS FROM ACOUSTIC REVERBERATION

Hafiz Malik

Department of Electrical and Computer Engineering
220 Engineering Lab Building (ELB)
University of Michigan-Dearborn
Dearborn, MI 48128

Hany Farid[*]

Department of Computer Science
6211 Sudikoff Lab
Dartmouth College
Hanover, NH 03755

ABSTRACT

An audio recording is subject to a number of possible distortions and artifacts. For example, the persistence of sound, due to multiple reflections from various surfaces in a room, causes temporal and spectral smearing of the recorded sound. This distortion is referred to as audio reverberation time. We describe a technique to model and estimate the amount of reverberation in an audio recording. Because reverberation depends on the shape and composition of a room, differences in the estimated reverberation can be used in a forensic and ballistic setting.

***Index Terms*—** Audio Forensics

1. INTRODUCTION

The past few years have seen significant advances in image forensics [1]. At the same time, techniques for authenticating audio recordings are less developed. Notable exceptions include a technique for classifying audio environments from a set of low-level statistical features [2], a technique that employs spectral distances and phase shifts [3], and the electric network frequency (ENF) criterion which verifies integrity by comparing the extracted ENF with a reference frequency [4].

Here we exploit specific artifacts introduced at the time of recording to authenticate an audio recording. Audio reverberation is caused by the persistence of sound after the source has terminated. This persistence is due to the multiple reflections from various surfaces in a room. As such, differences in a room's geometry and composition will lead to different amounts of reverberation time. There is a significant literature on modeling and estimating audio reverberation (see, for example, [5]). We describe how to model and estimate audio reverberation – this approach is a variant of that described in [6]. We show that reverberation can be reliably estimated and show its efficacy in simulated and recorded speech.

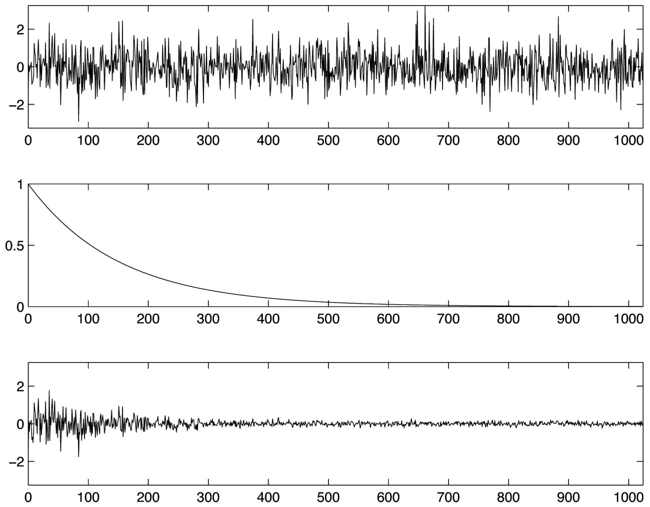

Fig. 1. Shown from top to bottom are: a signal $x(t)$; the exponential decay $d(t)$; and the resulting decayed signal $y(t)$ with additive noise.

2. METHODS

The decay of an audio signal $x(t)$ is modeled with a multiplicative decay and additive noise (Fig. 1):

$$y(t) \quad = \quad d(t)x(t) + n(t), \tag{1}$$

where,

$$d(t) \quad = \quad \exp(-t/\tau). \tag{2}$$

The decay parameter τ embodies the extent of the reverberation, and can be estimated using a maximum likelihood estimator.

We assume that the signal $x(t)$ is a sequence of N independently and identically-distributed (*iid*) zero mean and normally distributed random variables. We also assume that this signal is uncorrelated to the noise $n(t)$ which is also a sequence of N *iid* zero mean and normally distributed random variables with variance σ_n. With these assumptions, the

[*] This work was supported by a gift from Adobe Systems, Inc., a gift from Microsoft, Inc. and a grant from the National Science Foundation (CNS-0708209).

observed signal $y(t)$ is a random variable with a probability density function given by:

$$P_{y(t)}(k) \;=\; \frac{1}{\sqrt{2\pi\sigma^2\gamma^2(t)}}\cdot\exp\left(-\frac{k^2}{2\sigma^2\gamma^2(t)}\right), \quad (3)$$

where

$$\gamma(t) \;=\; \sqrt{\exp(-2t/\tau)+\sigma_n^2}. \quad (4)$$

The likelihood function is then given by:

$$L(y,\sigma,\gamma) \;=\; \frac{1}{(2\pi\sigma^2)^{N/2}\prod_{k=0}^{N-1}\gamma(k)}\cdot \exp\left(-\frac{1}{2\sigma^2}\sum_{k=0}^{N-1}\frac{y^2(k)}{\gamma^2(k)}\right). (5)$$

The log-likelihood function, $\ln(L(\cdot))$, is:

$$\mathcal{L}(y,\sigma,\gamma) \;=\; -\frac{N}{2}\ln(2\pi\sigma^2)-\sum_{k=0}^{N-1}\ln(\gamma(k)) - \frac{1}{2\sigma^2}\sum_{k=0}^{N-1}\frac{y^2(k)}{\gamma^2(k)}. \quad (6)$$

The decay parameter τ is estimated by maximizing the log-likelihood function $\mathcal{L}(\cdot)$. This is achieved by setting the partial derivatives of $\mathcal{L}(\cdot)$ equal to zero and solving for the desired τ.

$$\frac{\partial\mathcal{L}}{\partial\sigma} \;=\; -\frac{N}{\sigma}+\frac{1}{\sigma^3}\sum_{k=0}^{N-1}\frac{y^2(k)}{\gamma^2(k)} \quad (7)$$

$$\frac{\partial\mathcal{L}}{\partial\tilde{\tau}} \;=\; -\sum_{k=0}^{N-1}\frac{k\tilde{\tau}^{2k-1}}{\gamma^2(k)}\left(\frac{y^2(k)}{\sigma^2\gamma^2(k)}-1\right). \quad (8)$$

For the purpose of numerical stability, the maximization is performed on $\tilde{\tau}=\exp(-1/\tau)$. Although σ in Equation (7) can be solved for analytically, $\tilde{\tau}$ in Equation (8) cannot. As such, an iterative non-linear minimization is required. This minimization consists of two primary steps, one to estimate σ and one to estimate $\tilde{\tau}$. In the first step σ is estimated by setting the partial derivative in Equation (7) equal to zero and solving for σ, to yield:

$$\sigma^2 \;=\; \frac{1}{N}\sum_{k=0}^{N-1}\frac{y^2(k)}{\gamma^2(k)} \;=\; \frac{1}{N}\sum_{k=0}^{N-1}\frac{y^2(k)}{\tilde{\tau}^{2k}+\sigma_n^2}. \quad (9)$$

This solution requires an estimate of σ_n, which is estimated from the noise floor following the decayed signal. This solution also requires an estimate of $\tilde{\tau}$ which is initially estimated using Schroeder's integration method [9]. In the second step, $\tilde{\tau}$ is estimated by maximizing the log-likelihood function $\mathcal{L}(\cdot)$ in Equation (6). This is performed using a standard gradient descent optimization, where the derivative of the objective function is given by Equation (8). These two steps are iteratively executed until the differences between consecutive estimates of σ and $\tilde{\tau}$ are less than a specified threshold. In practice, this optimization is quite efficient, converging after only a few iterations.

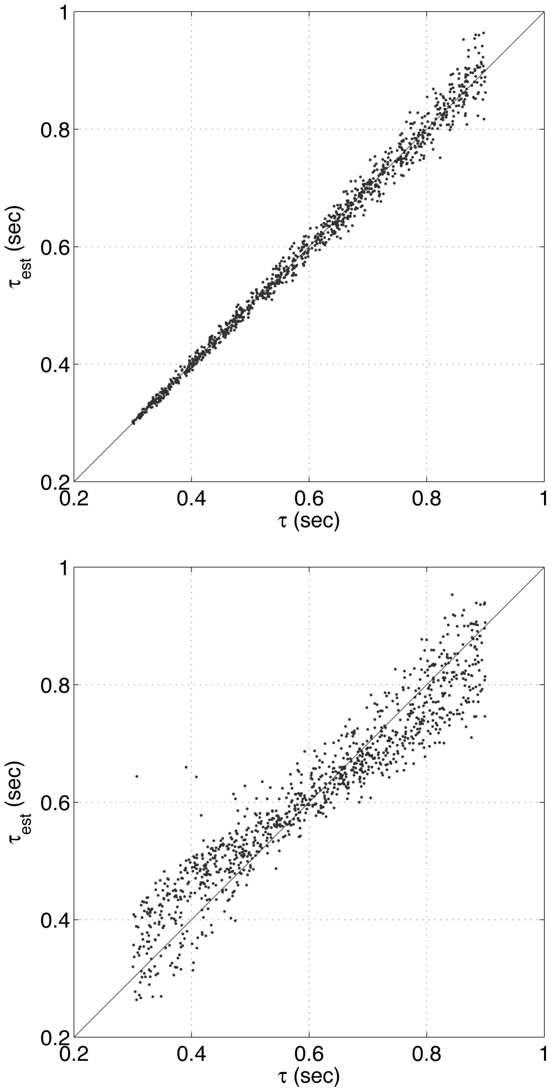

Fig. 2. Estimation results for synthetically generated signals with no noise (top) and with 26dB of additive noise (bottom).

3. RESULTS

Shown in the top panel of Fig. 1 is a signal $x(t)$ generated according to our *iid* zero mean and normally distributed assumption with $N = 1024$ and with an assumed sampling rate of 512 samples/seconds. Shown in the central panel is the exponential decay $d(t) = \exp(-t/\tau)$ with $\tau = 0.29$ seconds, and shown in the bottom panel is the resulting decayed signal $y(t)$ with additive noise as specified by Equation (1). We generated 1000 random signals according to this model with values of $\tau \in [0.29, 0.88]$ seconds, and either with no noise ($\sigma_n = 0$), or with a σ_n to yield an average signal-to-noise ratio of 26dB. As described in the previous section, the decay parameter τ was estimated from these signals. Shown in the top panel of Fig. 2 are the actual values of τ as a function of the estimated values (τ_{est}) for the no noise case. The average estimation error is 0.01 seconds with a standard deviation of 0.01. Shown in the bottom panel of Fig. 2 are the estimation results for the additive noise case. The average estimation error is 0.04 seconds with a standard deviation of 0.03. The handful of outliers have small values of τ (i.e., rapid decay) which leads to a signal where the noise dominates, thus leading to occasionally unreliable estimates.

In our second experiment we generated audio recordings with different amounts of reverberation using the model of [10]. Each recording was 9 seconds in length, and with a reverberation time of either $\tau = 0.3$ or $\tau = 0.6$ seconds. Each recording was corrupted with additive white noise with a signal to noise ratio of 35dB. We then created hybrid recordings with the first half having one reverberation time and the second having another. Because the underlying audio recordings were identical, there was no audible splice where the recordings were combined. As described above, the reverberation was estimated from eight positions, each of which were manually selected on the basis that the speech at these positions decayed to the noise floor. In the first example, the reverberation in the first half of the audio was 0.3 seconds, and in the second half it was 0.6 seconds. The mean (and standard deviation) estimate for the decay parameter τ for the first half is 0.062 (0.013) and for the second half is 0.083 (0.005). In the second example, the reverberation in the first half of the audio was 0.6 seconds, and in the second half it was 0.3 seconds. The mean (and standard deviation) estimate for the decay parameter τ for the first half is 0.088 (0.011) and for the second half is 0.052 (0.011). In each case, there was a significant difference in the estimated decay parameters, which could subsequently be used as evidence of manipulation.

In our third experiment, we recorded human speech in four different environments: (1) outdoors; (2) small office (7’ × 11’ × 9’); (3) large office (15’ × 11’ × 9’); and (4) stairwell. In each case, the same speaker read the opening paragraph of Charles Dickens’ *Tale of Two Cities*. The audio was recorded using a commercial-grade microphone. As described above, the reverberation was estimated from fourteen positions in each of the recorded audio segments. These were manually selected on the basis that the speech at these positions decayed to the noise floor, Fig. 4. Because there was considerable background noise in these recordings, each recording was initially pre-processed with a speech enhancement filter [11]. The mean (and standard deviation) estimate for the decay parameter τ, in seconds, is: (1) outdoors: 0.049 (0.013); (2) small office: 0.062 (0.017); (3) large office: 0.083 (0.012); and (4) stairwell 0.203 (0.064). This difference is significant as confirmed by a one-way ANOVA ($F(3, 40) = 39.93$, $p \leq 0.000001$). Although individual estimates of τ are not sufficiently reliable to fully characterize a speaker’s environment, the running averages over even a short length of audio shows significant differences in the estimated decay parameter.

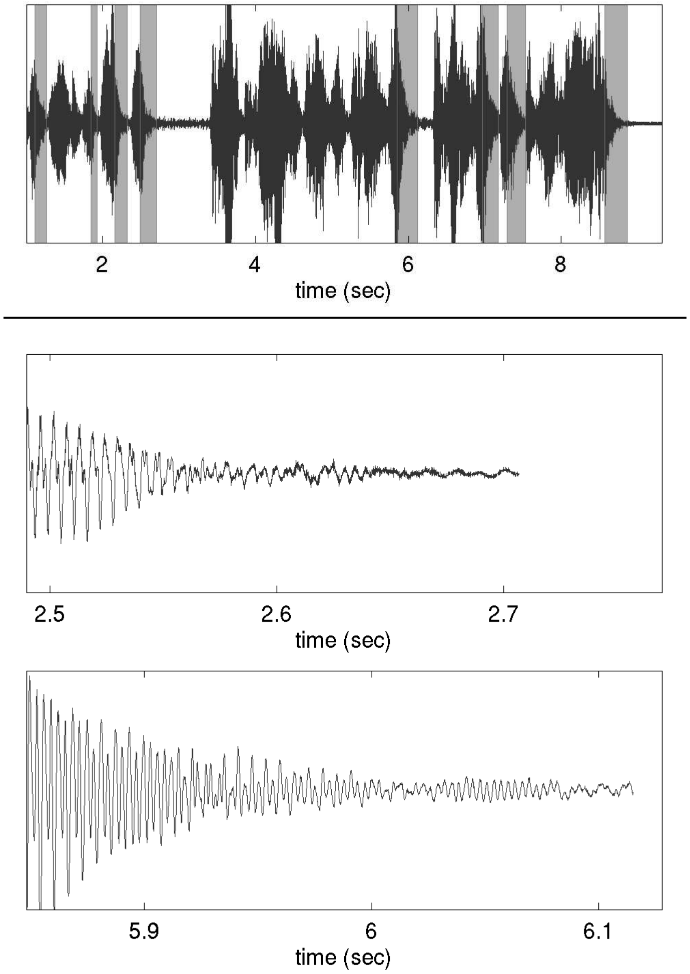

Fig. 3. Shown in the top panel is an audio signal whose left and right halves have different amounts of reverberation. The reverberation time was estimated from eight positions (shaded areas). Shown below are two sample segments from the left and right halves of the signal.

4. DISCUSSION

We have described how audio reverberation can be modeled, estimated, and used in a forensic setting. We have shown the efficacy of this approach on synthetically generated and recorded audio. We expect this approach to be a useful forensic tool when used in conjunction with other techniques that measure microphone characteristics, background noise, and compression artifacts.

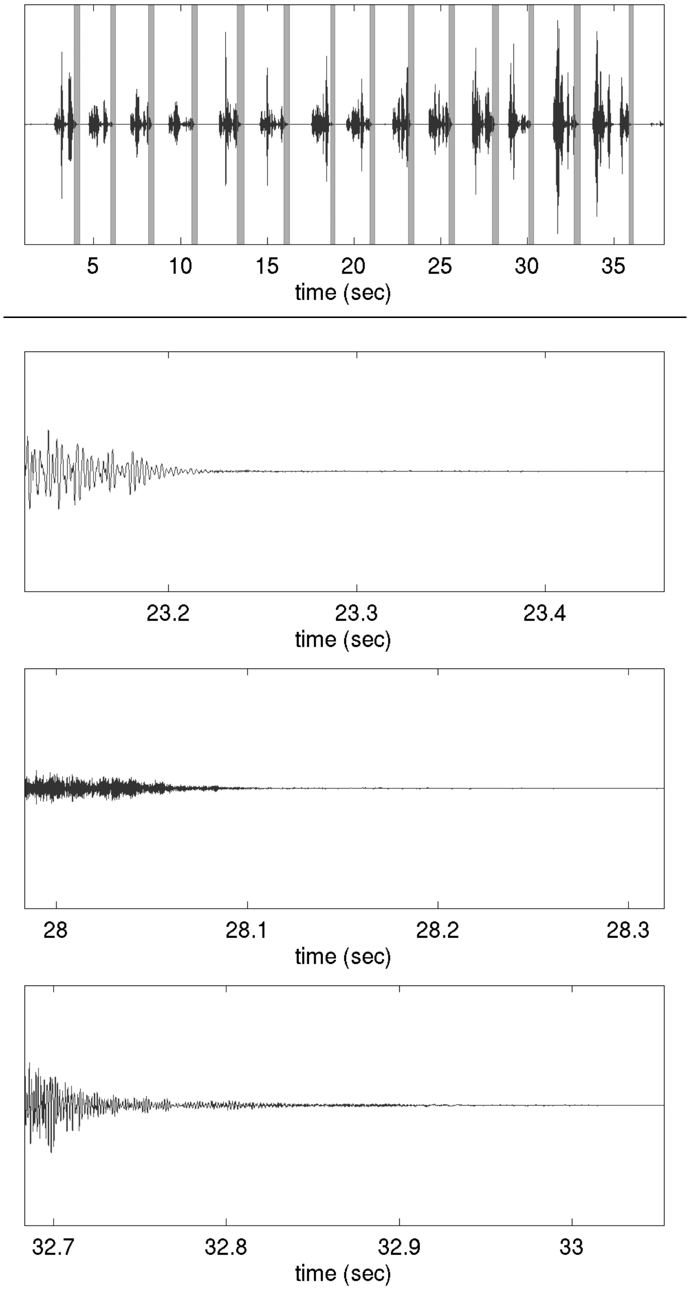

Fig. 4. Shown in the top panel is an audio signal recorded in a large office. The reverberation time was estimated from fourteen positions (shaded areas), each manually selected such that the speech decayed to the noise floor. Shown below are three sample segments revealing the form of the audio decay due to reverberation.

5. REFERENCES

[1] H. Farid, "A survey of image forgery detection," *IEEE Signal Processing Magazine*, vol. 2, no. 26, pp. 16–25, 2009.

[2] C. Kraetzer, A. Oermann, J. Dittmann, and A. Lang, "Digital audio forensics: a first practical evaluation on microphone and environment classification," in *Proceedings of the 9th workshop on Multimedia and Security*, Dallas, TX, 2007.

[3] D.P. Nicolalde and J.A. Apolinario, "Evaluating digital audio authenticity with spectral distances and ENF phase change," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, 2009.

[4] C. Grigoras, "Applications of ENF criterion in forensic audio, video, computer and telecommunication analysis," *Forensic Science International*, , no. 167.

[5] R. Ratnam, D.L. Jones, B.C. Wheeler, W.D. Obrien Jr., C.R. Lansing, and A.S. Feng, "Blind estimation of reverberation time," *Journal of Acoustic Society of America*, vol. 5, no. 114, pp. 2877–2892, 2003.

[6] H.W. Lollmann and P. Vary, "Estimation of the reverberation time in noisy environments," in *Proceedings of International Workshop on Acoustic Echo and Noise Control*, Seattle, WA, 2008.

[7] R.C. Maher, "Acoustical characterization of gunshots," in *Signal Processing Applications for Public Security and Forensics*, Washington, DC, 2007.

[8] G. Defrance, L. Daudet, and J.-D Polack, "Characterizing sound sources from room-acoustical measurements," in *International Symposium on Room Acoustics*, Oslo, Norway, 2008.

[9] M.R. Schroeder, "New method for measuring reverberation time," *Journal of Acoustic Society of America*, vol. 3, no. 37, pp. 409–412, 1965.

[10] E. Lehmann and A. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *Journal of Acoustic Society of America*, vol. 1, no. 121, pp. 269–277, 2008.

[11] Y. Lu and P.Loizou, "A geometric approach to spectral subtraction," *Speech Communication*, , no. 50, pp. 453–466, 2008.